DTW-based Phonetic Alignment Using Multiple Acoustic Features

Sérgio Paulo, Luís C. Oliveira

L^2F Spoken Language Systems Lab.
INESC-ID/IST
Rua Alves Redol 9, 1000-029 Lisbon, Portugal
{spaulo,lco}@l2f.inesc-id.pt

Abstract

This paper presents the results of our effort in improving the accuracy of a DTW-based automatic phonetic aligner. The adopted model assumes that the phonetic segment sequence is already known and so the goal is only to align the spoken utterance with a reference synthetic signal produced by waveform concatenation without prosodic modifications. Instead of using a single acoustic measure to compute the alignment cost function, our strategy uses a combination of acoustic features depending on the pair of phonetic segment classes being aligned. The results show that this strategy considerably reduces the segment boundary location errors, even when aligning synthetic and natural speech signals of different gender speakers.

1. Introduction

Phonetic alignment plays an important role in speech research. It is needed in a wide range of applications, from the creation of prosodically labelled databases for research into natural prosody generation, to the creation of training data for speech recognizers. Furthermore, the development of many corpus-based speech synthesizers ([1], [2]) requires large amounts of annotated data. Manual phonetic alignment of speech signals is an arduous and time consuming task. Thus, the size of the speech databases that can be labelled this way are obviously very limited, and the creation of large speech inventories requires some sort of automatic method to perform the phonetic alignment. While building a system to automatically align a set of utterances, two different problems can be found. First, we have to know the sequence of phonetic segments observed in those utterances. Then, we need locate the segment boundaries. The sequence of segments can be obtained by using a pronunciation dictionary or by applying a set of pronunciation rules to the orthographic transcription of the utterances. However, it is, usually, not possible to predict the exact sequence uttered by the speaker and we must take into account possible disfluencies, elisions, allophonic variations, etc. A technique to handle this problem was presented in [3]. In this work, we will assume that we already have the correct sequence of segments and we will focus on the task of locating the segment boundaries. Several approaches have been taken to try to locate these boundaries. The most widely used technique is the use of HMM-based speech recognizers (sometimes hybrid systems, based on HMM and Artificial Neural Networks) in forced alignment mode. This approach relies on the use of phone models built under the HMM framework. But many times, no such speech recognizer is available. However, at least for those which are working on speech synthesis, speech synthesizers are available, and the phonetic alignment can be also performed automatically using the speech synthesis based approach ([4],[5]). This approach starts by producing a synthetic speech signal with the desired phonetic sequence allowing us to know the exact location of the phonetic segment boundaries. The next step is to compute, every few milliseconds, vectors of acoustic features for both the synthetic and natural speech signals. By using some type of distance measure, the acoustic feature vectors can be aligned with each other using the DTW algorithm ([6]). The result is a time alignment path between the synthetic and natural signal time scales, that allows us to map the segment boundaries from the synthetic signal into the natural utterance. This approach does not require any previously segmented speech from the same speaker but the results depend, in some extent, on the similarity between the synthesizer's and speaker's voice. However, our results show that the performance of this method is strongly dependent on the selection of the acoustic features used for computing the cost function of the alignment procedure.

In this paper we will try to improve the performance of a DTW-based phonetic aligner by combining multiple acoustic features depending on the class of segments being aligned. In section 2, we will start by describing the process to produce the synthetic reference signal. The next section (section 3) presents the automatic procedure used for the selection of the most discriminative acoustic features for each pair of segment classes. The computation of the alignment cost and the alignment procedure is described in section 4. The results of the proposed method as well as the ones produced by a traditional method are presented in section 5, where we also describe some efforts on the reduction of the phonetic boundary errors when the synthesized and the recorded speech signals belong to speakers of different gender.

2. Synthesized speech signal

The speech synthesis based phonetic alignment relies on a synthesized speech signal that is used as a reference for the alignment procedure. This signal can be produced by using some sort of a speech synthesizer, that can be modified to produce the desired phonetic sequence together with the segment boundaries. However, this solution as the disadvantage that the synthetic signal may have undesirable distortions introduced by the signal processing required to modify its rhythm and intonation. For our purposes, these prosodic modifications are not necessary and a simple waveform concatenation system was used. Since our goal is to locate the segment boundaries, we used diphones as concatenation units so that the possible concatenation discontinuities in the middle of the phone do not affect the location of the phone boundary. To perform this concatenation we have modified the unit selection module of the Festival Speech Synthesis System ([7]). A local search around the diphone boundaries tries to find the best concatenation point. The Inverse Harmonic Mean distance (IHM,[8]) between the line

spectral frequencies (LSFs) was used to evaluate the spectral discontinuity. This distance measure can be expressed as:

$$d_{IHM}(x_i, y_j) = \sum_{p=1}^{P} c(p)(x_i(p) - y_j(p))^2 \quad (1)$$

where $x_i(p)$ is the p^{th} LSF coefficient of the first unit, $y_i(p)$ is the p^{th} LSF coefficient of the second unit. The weighting factor, $c(p)$, is defined as:

$$c(p) = \left[\frac{1}{\omega(p) - \omega(p-1)} + \frac{1}{\omega(p+1) - \omega(p)}\right] \quad (2)$$

with $\omega(0) = 0$, $\omega(P+1) = \pi$ and $\omega(p) = x_i(p)$ or $\omega(p) = y_j(p)$ so that $c(p)$ is maximized. Intuitively, since this weighting is based on neighboring LSFs, mismatch in spectral peaks is weighted more heavily than mismatch in spectral valleys. This results from the LSF parameters property that closely spaced LSFs are indicative of resonances in the spectrum while more widely spaced LSFs indicate a flat-shaped spectral structure.

3. Acoustic Features

For the selection of the alignment cost function, we considered some of the most relevant acoustic features used in speech processing: the Mel-frequency cepstrum coefficients (MFCC) and their differences (deltas), the four lowest resonances of the vocal tract (formants), the line spectral frequencies (LSF), the energy and its delta and the zero crossing rate of the speech signal. Both the energy and the MFCC coefficients, as well as their deltas, were computed using software from the Edinburgh Speech Tools Library[9]. The formants were computed using the *formant* program of the Entropic Speech Tools[10] and the remaining features were computed using our own programs. Our first experiments showed that each of these features used separately produced uneven results. Depending on the class of phones to be aligned some features proved better than others. For instance, in a vowel-plosive transition, the energy feature was the best performer, but for vowel-vowel transition, the best results were achieved using formants as features. This immediately suggested the use of multiple features to distinguish the different phone transition classes.

3.1. Feature normalization

The combination of multiple features requires some form of normalization to equalize their influence on the overall alignment cost. It was decided to normalize the values to the range $[0, 1]$.

The first stage was to find which features had values with a *Gaussian* distribution. Observing the histograms of each coefficient, the MFCCs and their deltas were the only ones that could match that distribution. The mean and standard deviation were computed for each one of them, and the normalization was then performed using the equation:

$$x_i = \frac{1}{2} + \frac{X_i - \mu_i}{4\sigma_i} \quad (3)$$

where x_i, X_i, μ_i and σ_i are the normalized value, the non-normalized value, the mean value, and the standard deviation of the i^{th} MFCC, respectively. The LSF values were divided by π. Since the zero crossing rate was computed by evaluating the ratio between the number of times the speech signal crosses the zero magnitude and the number of signal samples existing in a fixed size window (some milliseconds), its values have already the right magnitude (between 0 and 1). For the energy, its delta and for the formants, maximum and minimum values were found for each utterance, and mean values were computed for each of them ($\overline{Y_{i_{max}}}$ and $\overline{Y_{i_{min}}}$). The normalized values were calculated using the following equation:

$$y_i = \frac{Y_i - \overline{Y_{i_{min}}}}{\overline{Y_{i_{max}}} - \overline{Y_{i_{min}}}} \quad (4)$$

3.2. Feature Selection Procedure

The next step was to find the most discriminating features in a given phonetic context. That is, which feature allowed us to locate the boundary with greater precision. For this purpose we had a set of 300 manually aligned utterances that were used to evaluate the relevance of each feature. These utterances were spoken by a different speaker than the one used to record the diphone inventory (although both of them were male). The speech synthesizer was then used to produce reference synthetic signals for the phonetic sequences of these utterances and sets of feature vectors were computed every 5 milliseconds for both the reference and spoken signals. A distance matrix between the two sequences was then computed for each feature. Fig. 1 shows a rough representation of this matrix. We then evaluated each feature on its capacity to discriminate the difference between two consecutive phones. This was achieved by computing the average distance between feature vectors of the same phone ($\overline{dist_s}$), and of different phones ($\overline{dist_d}$). Using the example in Fig. 1, if we want to choose an acoustic feature to distinguish the *silence* (#) and the vowel *u*, the $\overline{dist_s}$ is the average of the values in regions 1 and 6 on that matrix, while the $\overline{dist_d}$ is the average of the values on regions 2 and 5.

This procedure was performed for every pair of phones and for every utterance on the training set, and its resulting values were saved at the end of each iteration. Finally, we computed an average value of the ratio between $\overline{dist_s}$ and $\overline{dist_d}$ for each pair of phonetic segments and for each acoustic feature. The chosen feature is the one that gives a minimal value for this ratio using the equation:

$$F_k = \min_x \sum_{i=1}^{N_k} \frac{\overline{dist_s}(k, x, i)}{\overline{dist_d}(k, x, i)} \quad (5)$$

where, x is one of the tested features, k represents the pair of phones that is being analyzed, N_k is the number of instances of this pair in our set of utterances, F_k is the best feature for this type of transition, and $\overline{dist_s}(k, x, i)$ and $\overline{dist_d}(k, x, i)$ are the mean distances for the instance i using the acoustic feature x. The smaller is that ratio, the greater is probability of having well aligned frames, at least locally. With this approach, we are trying to use the feature that assigns the greatest penalty for the alignment paths when they fall out of the darkest regions of the Fig. 1 (regions 1, 6, 11 and 16).

Given the reduced amount of training data, we soon realized that it would be impossible to have a large enough number of instances, for each pair of segments to produce reliable results. Thus, the different phonetic segments were grouped into phonetic classes: vowels, fricatives, plosives, nasals, liquids and silence. The semi-vowels were included in the vowel class. The described procedure for differentiating the phones was then repeated using phone class transitions (vowel-vowel, fricative-vowel, etc.).

The analysis of the results showed that, in general, for each phone class transition, at least two of the tested features showed good discriminative capacity. This could suggest some equivalence between the two features but it could also mean that the

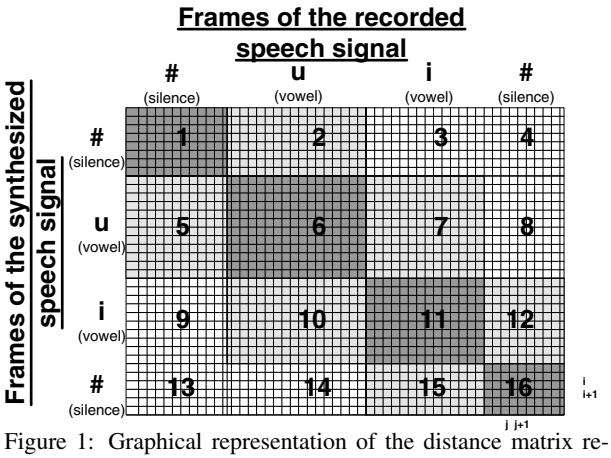

Figure 1: Graphical representation of the distance matrix regions used for choosing the best feature / pair of features to align the different pairs of phonetic segments.

	N	F	L	P	#	V
N	frm+lsf	mfcc+zcrs	frm+en	lsf+en	frm+en	mfcc+mfcc
F	lsf+lsf	mfcc+en	en+zcrs	lsf+en	zcrs+en	lsf+lsf
L	lsf+en	lsf+en	lsf+lsf	mfcc+en	mfcc+en	frm+mfcc
P	lsf+en	lsf+lsf	lsf+en	mfcc+mfcc	lsf+zcrs	mfcc+en
S	lsf+en	lsf+en	lsf+en	lsf+en	x	lsf+en
V	mfcc+en	zcrs+lsf	mfcc+en	lsf+en	mfcc+en	frm+mfcc

Table 1: Best feature pairs for the multiple phonetic segment class transitions.

two features were complementary. This way we performed a combined optimization to select the pair of features for each phone class pair. The process could be extended to a combination of even more features but the results showed that there was no significant improvement in using more than two features simultaneously.

3.3. Selected features

The Table 1 shows the results of this procedure, where N, P, F, V and L symbols are the classes of the nasal, plosive, fricative, vowel and liquid phonetic segments, respectively. The # symbol is the silence. In the same table, mfcc, lsf, frm, en and zcrs are the MFCC coefficients and their deltas, LSFs, formants, energy and its delta and the zero crossing rate, respectively. The x symbol means that this class transition does not exist in our training set. The best feature pair for a transition x-y, is located on the line of x and column of y.

4. Frame Alignment

Before applying the DTW algorithm the distance matrix between the reference and the spoken signal must be built, we used the Euclidean distance for this purpose. Since we know the boundary locations of the synthetic segments, the distance matrix can be built iteratively, phone-pair by phone-pair.

Taking the example shown in Fig. 1, to build the distance matrix we start by computing the matrix values for all the rows that correspond to the phone-pair *#-u* using the best pair of features, based on the former results. However, the phone *u* also belongs to the next phone-pair (*u-i*) and the computed distance is multiplied by a decreasing triangular weighting window. The distance for the next phone pair (*u-i*) is then computed using the best pair of features for the vowel-vowel transition and its value is added to the rows corresponding to segment *u* weighted by an increasing triangular window. Fig. 2 shows this weighting

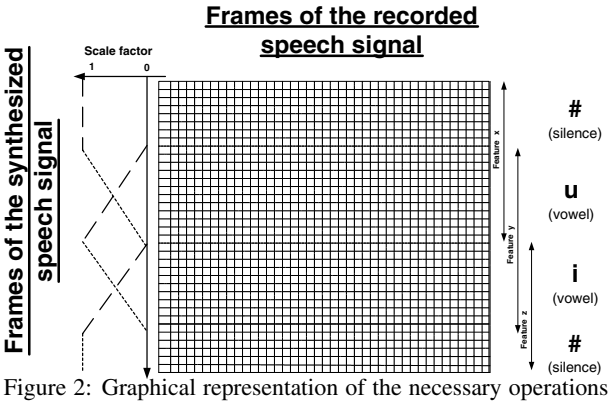

Figure 2: Graphical representation of the necessary operations for building the distance matrix.

triangular windows, where the dotted lines are the weighting factor of the previous phone-pair distances and the dashed lines are the weights of the distances for the next phone-pair. After computing all the values of the distance matrix, the DTW algorithm is applied to find the path that links the top left corner of the matrix to the lower right corner with a minimum accumulated distance. This path will be the alignment function between the time scale of the synthetic reference signal and the spoken utterance.

5. Evaluation

The procedure described in the previous section was applied in two different experiments:

- Alignment of the utterances included in the training set;
- Alignment of another speech database utterances, where the synthetic and natural voices belong to a male and a female speakers, respectively.

5.1. Training set utterances

The results of this experience are depicted in Fig. 3, where the dotted line is the annotation accuracy when the entire set is aligned using a single acoustic feature consisting of a vector of 12 Mel-frequency cepstrum coefficients and their differences. Only 46% of the phonetic segments were aligned with an error less than 20 ms. Using only the best feature for computing the distance of each phone class pair increases the 20ms accuracy to 59% of the segments (dashed line). This result can be improved to 70% by combining two features for computing the distance measure (solid line).

The relatively low percentage of agreement for tolerances lower than 20ms can be partially explained by the fact that the segmentation criteria used in the annotation of the reference corpus was not exactly the same as the one used in the segmentation of the logathomes used to produce the synthetic reference. Another difficulty was that the speech material in the reference corpus was uttered by a professional speaker with a very rich prosody and large variations in energy, where several consecutive voiced speech segments become unvoiced. This is, in our opinion the main reason for about 4% of disagreement within high tolerances (about 100 milliseconds).

5.2. Speakers of different gender

We also tried to align a speech database recorded by a female speaker using the proposed method, but for this female voice the formant detection tool did not generate values with an high enough accuracy. This was not a surprise, since the detection of the spectral envelope is more difficult for high-pitched voices. This way, we replaced the feature pairs that use formants by the

	N	F	L	P	#	V
N	mfcc+lsf	mfcc+zcrs	mfcc+en	lsf+en	mfcc+en	mfcc+mfcc
F	lsf+lsf	mfcc+en	en+zcrs	lsf+en	zcrs+en	lsf+lsf
L	lsf+en	lsf+en	lsf+lsf	mfcc+en	mfcc+en	en+mfcc
P	lsf+en	lsf+lsf	lsf+en	mfcc+mfcc	lsf+zcrs	mfcc+en
S	lsf+en	lsf+en	lsf+en	lsf+en	x	lsf+en
V	mfcc+en	zcrs+lsf	mfcc+en	lsf+en	mfcc+en	frm+mfcc

Table 2: Best feature pairs for the multiple phonetic segment class transitions without the formants.

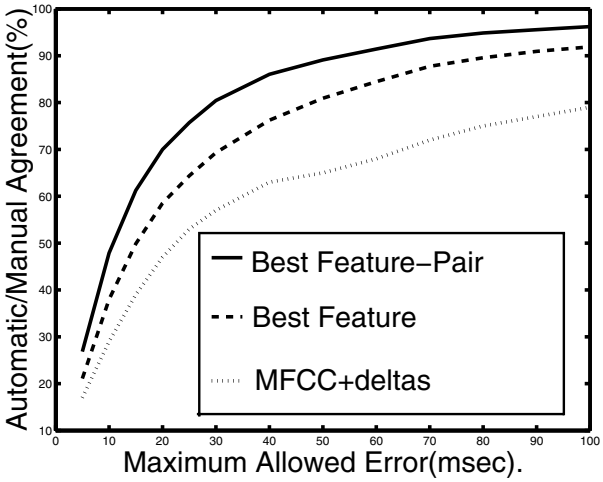

Figure 3: Accuracy of some versions of the proposed method and a classic speech synthesis based aligner.

second best feature pair determined during the feature selection step that did not include the formant values. The Table 2, shows the feature pairs that we used for this experiment. After this modification, the alignment was performed and the results are presented in Fig. 4, where the dotted line represents the accuracy of the classic DTW-based aligner once more, and the dashed line is the one we get with the proposed method. The solid line represents the accuracy of our method after applying a vocal tract length normalization technique, explained in [11], to the synthetic signal spectral features. This encouraging results shows the robustness of our alignment strategy, even when the involved voices are significantly different.

6. Conclusions

In this work we have presented a method for selecting the most relevant acoustic features for aligning two speech signals with the same phone sequence but with different durations. This features were then used in a DTW-based algorithm for performing the phonetic alignment of a spoken utterance. The results clearly show the advantage of using multiple acoustic features, selecting the most appropriate pair for each class of segments in the alignment of two utterances: the most commonly used feature, MFCCs, performed well bellow the proposed method. We also showed the robustness of the proposed method that can even produce good results when aligning utterances recorded by speakers of different gender. In this case the system accuracy can be improved by applying a vocal tract length normalization technique.

7. Acknowledgements

The authors would like to thank M. Céu Viana and H. Moniz for providing the manually aligned reference corpus. This work is part of Sérgio Paulo's PhD Thesis sponsored by a Portuguese Foundation for Science and Technology (FCT) scholarship. This work was partially funded by the FCT project "Contador de Histórias" of the POSI Program. INESC-ID Lisboa also had support from the POSI Program.

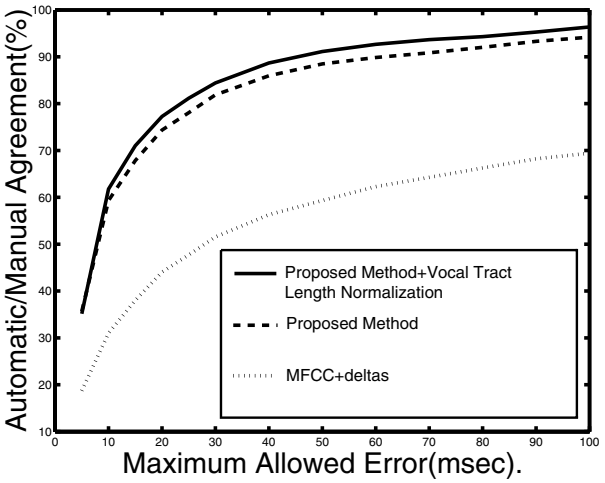

Figure 4: Accuracy of the proposed method when applied to utterances of a different gender speaker.

8. References

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal, *The AT&T Next-Gen TTS System*, 137^{th} Acoustical Society of America meeting, Berlin, Germany, 1999.

[2] A. Black, *CHATR, Version 0.8, a generic speech synthesizer*, System documentation, ATR-Interpreting Telecomunications Laboratories, Kyoto, Japan, 1996.

[3] S. Paulo and L. Oliveira, *Multilevel Annotation of Speech Signals Using Weighted Finite State Transducers*. In Proceedings of IEEE 2002 Workshop on Speech Synthesis, Santa Monica, California, 2002.

[4] F. Malfrère and T. Dutoit, *High-Quality Speech Synthesis for Phonetic Speech Segmentation*. In Proceedings of Eurospeech'97, Rhodes, Greece, 1997.

[5] N. Campbell, *Autolabelling Japanese TOBI*. In Proceedings of ICSLP'96, Philadelphia, USA, 1996.

[6] Sakoe H. and Chiba, *Dynamic programing algorithm optimization for spoken word recognition.* IEEE Trans. on ASSP, 26(1):43-49, 1978.

[7] A. Black, P. Taylor and R. Caley, *The Festival Speech Synthesis System.* System documentation Edition 1.4, for Festival Version 1.4.0, 17th June 1999.

[8] R. Laroia, N. Phambo and N. Farvardin, *Robust efficient quantization of speech LSP parameters using structured vector quantizers.* In Proceedings of ICASSP'91, Toronto, Canada, 1991.

[9] P. Taylor R. Caley, A. Black, S. King, *Edinburgh Speech Tools Library* System Documentation Edition 1.2, 15th June 1999.

[10] *ESPS Programs Version 5.3* Entropic Research Laboratories Inc., 1998.

[11] E. Eide and H. Gish, *A parametric approach to vocal tract length normalization.* In Proceedings of ICASSP'96, Atlanta, USA, 1996.